Improving decision making for shots

Benjamin Larrousse
benj.larrousse@gmail.com

Abstract

The idea behind this paper is to analyze decision making regarding shots in football games. As *Statsbomb* released new data regarding shots, it is now possible to know in details conditions in which shots are taken. For now, this photo of shots is not used in many applications despite its great potential. In the following we will use it in order to understand if shooting was the best option, considering all players position. Quantifying this can have a range of applications, from macro analysis of shots to more specific player by player intuition. We will give some applications here, after modeling "better options" regarding shots.

1. Introduction

Data usage in sports has been popularized by the book (and the movie) "*Moneyball: The Art of Winning an Unfair Game*" [Ref 1] by Michael Lewis (2003), which talks about how Billy Beane used statistics in Baseball to find a new way of evaluating players, leading to better decisions regarding the draft. It is now finding its way in Football (soccer), with clubs more and more eager to find winning edge through statistical analysis [Ref 6].

At the heart of the data revolution in Football lies a statistic called *Expected Goals* [Ref 2]. It's a measure of the quality of chances created and conceded in a football game. More specifically, it's the probability of scoring based on several characteristics of a shot. It's based on historical shots, and some factors taken into account are: distance from goal, angle of the shot, body part, type of assist, etc. See Figure 1 for an illustration of the concept.

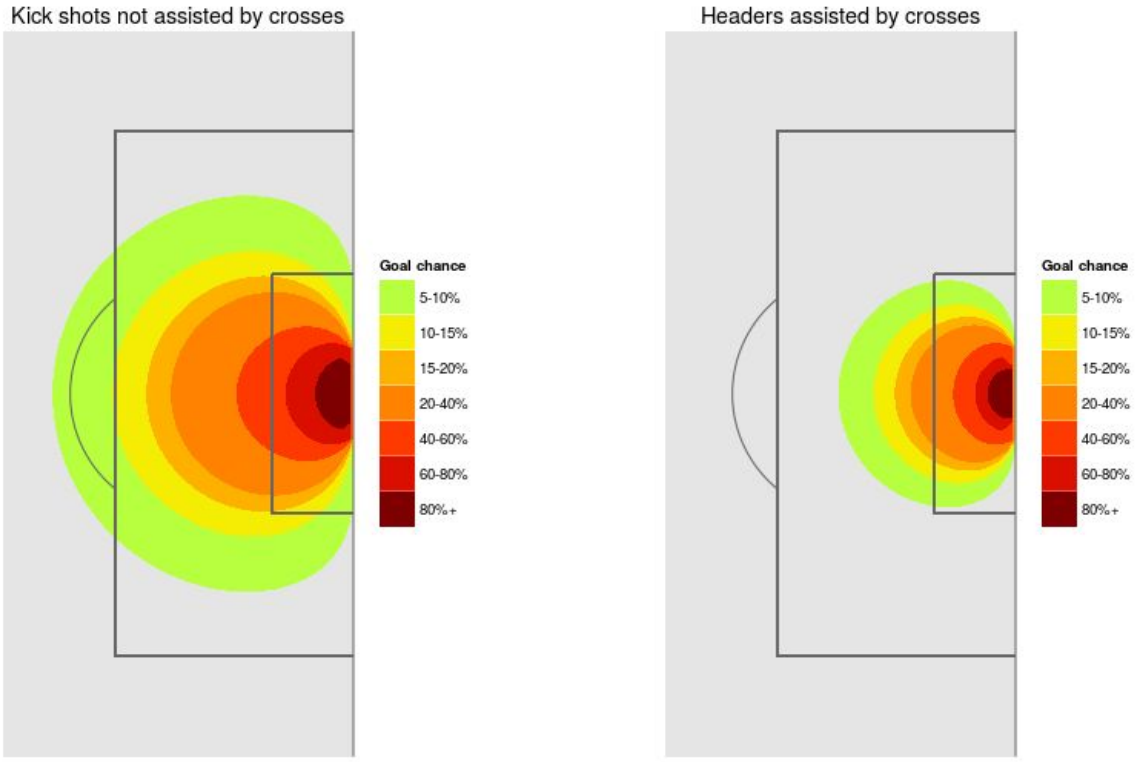

Fig. 1. Illustration of *Expected goals* (*xG*). Source: https://statsbomb.com/2016/04/explaining-and-training-shot-quality/

This metric, despite its usefulness, has some limitations[1]:

- it's not taking into account defensive pressure which we know influence shot quality a lot,
- it does not consider the goalkeeper's position.

This is why in 2018 *Statsbomb* released new data: ***freeze frame***. This data includes position of every player around the ball for every shot taken, goalkeeper included. See an example in Fig 2.

[1] For a full description, see: https://statsbomb.com/2018/05/statsbomb-data-launch-beyond-naive-xg/

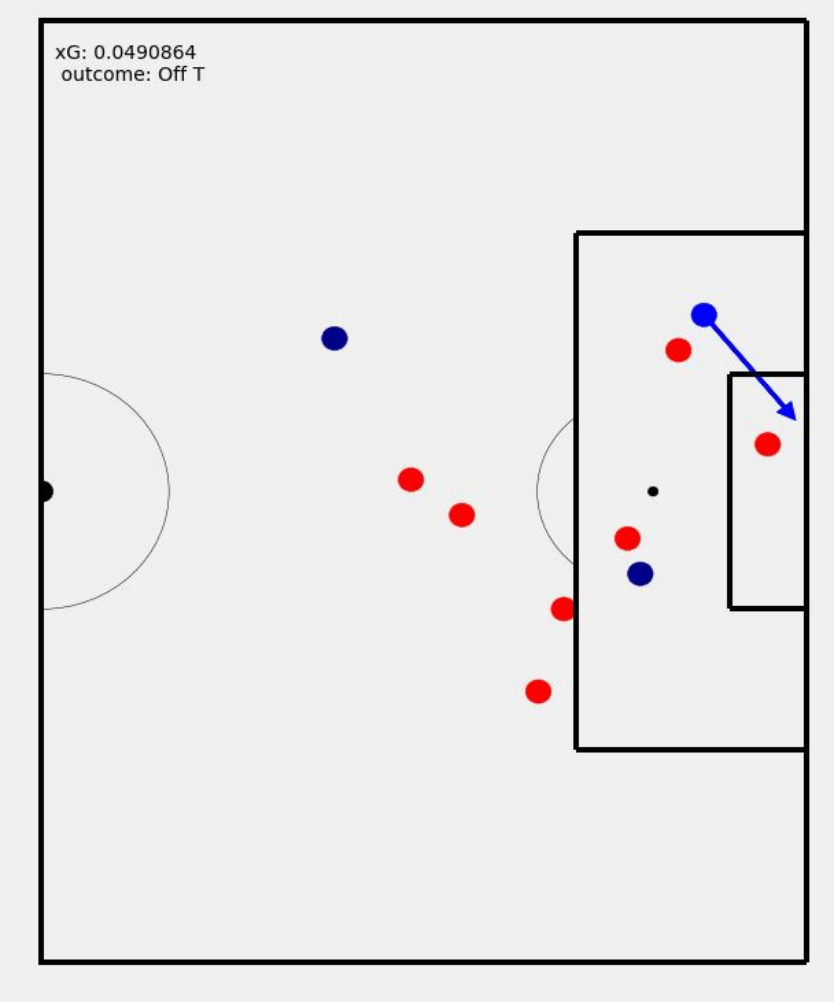

Fig 2: Data visualization of a freeze frame object (from *Statsbomb* data).

This allows to analyze shots in a more accurate way, and in particular it gives us an enhanced expected goals model. It will help for a number of use cases in Football, like analyzing defensive organisation or goalkeepers evaluation.

Whatever the goal (scouting, team analysis, player analysis, etc.), at the end the knowledge acquired from the data will serve one purpose: making smarter decisions. It could be finding undervalued players to buy, setting up tactics for the next game, or improving the general level of each player.

As freeze frame is relatively new, there is still very little use of it. One related paper is "*Beyond expected goals*" [Ref 3]. The general idea of the paper is to go beyond shots and evaluate the probability that a player without the ball will score, even in cases where the ball never reaches him. Authors create a model that gives this probability, based only on the instantaneous game state. The data used is tracking data, which gives them the opportunity to have velocity and acceleration of players. An information that we don't have with the freeze frame. Their *opportunity* model has many applications: post-match analysis, player-specific opportunity map to identify where players are the most dangerous, or scouting.

The aim of this paper is to explore how we can make smarter decisions in Football, specifically regarding one key event: shots. The *freeze frame* data allows us to analyse in detail the context in which shots are taken. We now know where players were when the shot was taken, thus it's possible to explore other options rather than shooting. Was there a better option? We will define below what we mean by "better" and how we can measure it. When a framework is fixed for the analysis, it will have a range of applications: it will be possible to analyze broadly how often shooting was the right decision. It could also be applied to making players better regarding shots, or for scouting. Applications will be discussed in Section Results and applications.

2. Data

For this analysis we use football game data provided by *Statsbomb* specifically for the Research competition of their first conference on Innovation in Football[2]. Games are from the French Ligue 1, 2017-2018 and 2018-2019 seasons. It's 760 matches, 2588566 events, 19328 shots, and 19070 freeze frames.

Each game is a json file containing match events. Events have a lot of characteristics: ID, period, type, team, player, location, etc. For a complete description, see: https://github.com/statsbomb/open-data. As the data is normalized, preprocessing is not a huge problem here. Data is processed with *Python* and stored in a Pandas dataframe. This way we can easily filter shot events and its characteristics (location, end location, outcome, type, technique, expected goal value, freeze frame, etc.).

Freeze frames contain location, name, and position of every player around the ball when a shot is taken (Fig. 2). It is used by *Statsbomb* to model xG, thus creating a model with more information taken into account. It will be the central data of our analysis.

3. Methodology

As explained before, we want to find if a player had other options (better ones) rather than taking a shot. By better, we mean making a pass to a teammate with a higher probability of scoring, a higher xG value. It can be measured by calculating the expected goal of a shot from this particular teammate.

Tor this, we need to be able to do two things:

- Calculate if a pass is possible from the shooter to a teammate,
- For teammates with a possible pass, calculate the xG after this pass, if the player takes a shot from his location. Of course, it's not mandatory that this particular player will shot directly,

[2] https://statsbomb.com/conference/

or shot at all. Maybe he will dribble, pass the ball to someone else, or lose it. But to start simple, we can assume options are evaluated based on a shot from the teammate's location.

It is the two main challenges that we have to face. In the following sections, we give more details about how we tackle them.

3.1 Expected goal calculation

In order to be able to compare the actual shot with an hypothetical shot from a teammate, we need to approximate the xG value for each teammate, depending on his location. The thing is, at *Statsbomb*, the freeze frame information is taken into account for the calculation of xG as it gives a more precise model. And of course, we don't have the freeze frame for an hypothetical shot, so it's not possible to calculate an xG value in the same way.

Nevertheless, it's possible to approximate an xG value based on location, by the mean or the median from similar shots. We choose to cut the field into zones and to calculate the mean and median in each zone. We only have two seasons data so it could be more precise (as the amount of data is relatively small) but it's a good starting point.

The size of zones we choose is 5 by 5. From the data we have, it's a good tradeoff between sufficient number of shots in each zone[3] and small enough zones. You can see the distribution in Fig. 3. Smaller zones could lead to better precision if we have (much) more data.

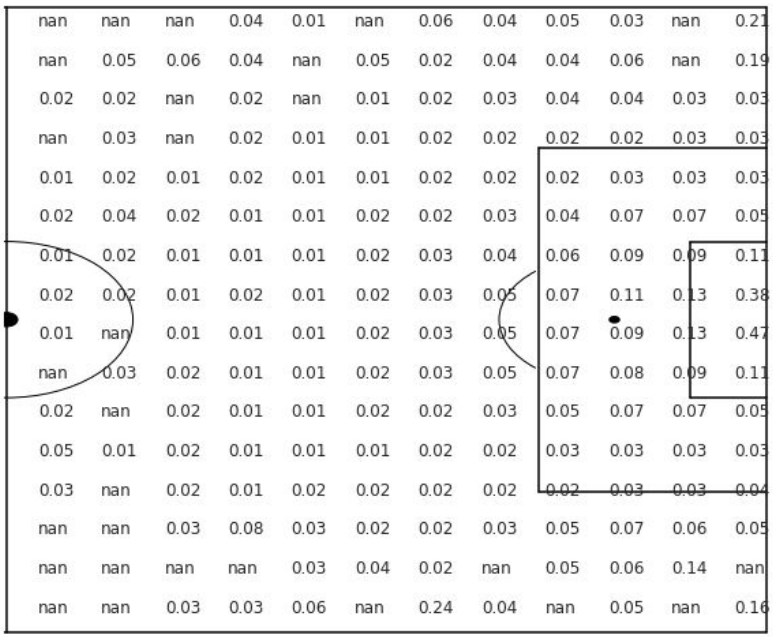

Fig. 3: xG value approximation based on statsbomb xG, cut by zones.

[3] Note that we're only looking at zones in which shots are generally taken

It should be noted that we're only using shots from open play (thus removing penalties and free-kicks).

In the following we choose to use the median and not the mean. They're not so far away from each other so it does not make a huge difference, but as median is most robust to outliers we think it's the best choice here. To illustrate further, we plot *statsbomb_xg* distribution together with the distribution of the median depending on zones (Fig 4 and Fig 5).

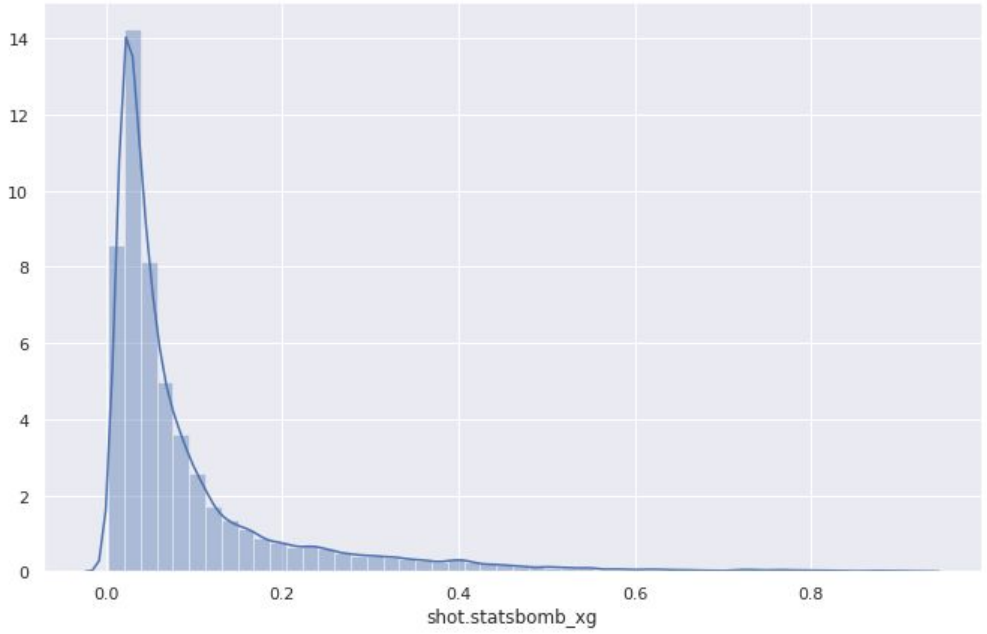

Fig. 4: *Statsbomb* xG value distribution (Open play shots only)

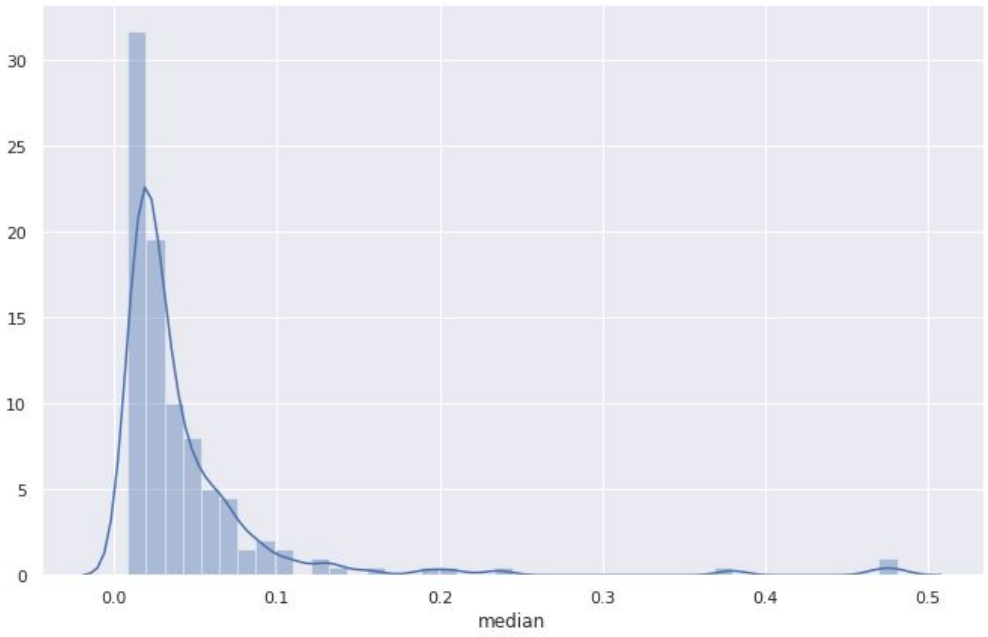

Fig. 5: median xG value by zones distribution (Open play shots only)

With the xG approximation, the next problem is to characterise possible passes.

3.2 Passing possibilities

We explore two ways of knowing if a pass is possible from the freeze frame. In any case, we will not consider off-target teammates, as a pass is not legally possible.

The first idea is to use *Voronoi diagrams* [Ref 4][Ref 5] to partition the pitch based on distances between players. A cell is associated to each player as illustrated in Fig. 6. We can know which teammate is connected to the shooter through his cell. We could say a pass is possible if a teammate's cell is connected to the shooter's cell because there is space to make a pass. But it's not enough, because we could miss some possible passes where cells are not connected but a pass is still possible.

We can see in Fig. 6 that a teammate is free on the left side of the box, but his Voronoi cell is not connected to the shooter so a pass to this particular teammate would be ignored by the model.

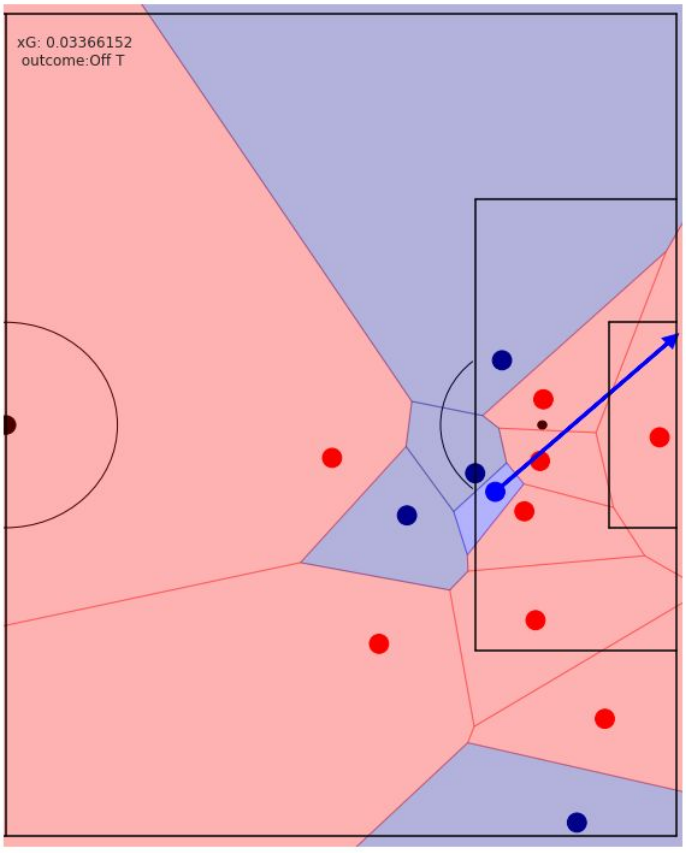

Fig. 6: Visualization of Voronoi cells on a freeze frame object Teammates are in blue, opponents in red.

We had to think of an other way of doing things. We want to know if an opponent is in the way between the shooter and a teammate. Mathematically, we can calculate if an opponent's location is in the segment between the shooter and a teammate. If so, a pass would not be possible. The problem is that in reality, a player is not a point with zero volume.

So what we need to look at is if a zone around an opponent's location is intersecting the shooter-teammate segment. This zone should correspond to a possible interception of the player. As the freeze frame does not give us the speed and direction of players (as opposed to tracking data), the best zone we can have is a circle. We choose to use a radius of 1.5 which seems empirically reasonable, but this value could be further improved.

This method leads to a separation between teammates with a possible pass and teammates without a possible pass.

3.3 Better options

Last but not least, now that we have a list of possible passes and an xG associated to each teammate's location, we need to know the probability that the pass will be successful. As we did for the xG values, we can calculate a mean probability from a pitch zone to any other pitch zones. This probability will be the percentage of completion of passes in our dataset.

We're not calculating a mean percentage of completion by player here. Averaging every pass from a player will not reflect his true ability on a particular zone (wing vs center, distance from goal, etc.) and a particular pass type (forward, backward, low range, mid range, etc.). As for the xG approximation, the amount of data we have is relatively small, and more data would lead to a more precise approximation.

At the end, passing the ball to a teammate *T* is a better option if:

- A pass is possible from the shooter to *T*,
- The pass probability multiplied by the estimated xG value of a shot from *T* is greater than the xG value of the actual shot.

In some cases there will be more than one better option, but If we only need one better option, we can only keep the one with the greatest xG value, or the one with the easiest pass.

In the next section we discuss our findings and some applications of this method.

4. Results and applications

This first analysis of decision making for shots can have many applications:

- From a broad perspective, we can see how often the decision to take a shot was the right one. Is this correlated to *xG* ? Do teams with a better ranking have less bad decisions?

- Players analysis: are players good at making decisions? Do they get better with time, from one season to the next (especially for young players still in development)? From a scouting perspective, it can add more information about the quality of a player. From a club's perspective, it can be useful to analyze their players' shots in order to help them make better decisions when similar situations happen again.

In the next subsections we will illustrate these applications.

4.1 Shot decisions VS xG value

First we can look at the percentage of shots were at least one better option exists, depending on *xG* value. This is represented in Fig. 7.

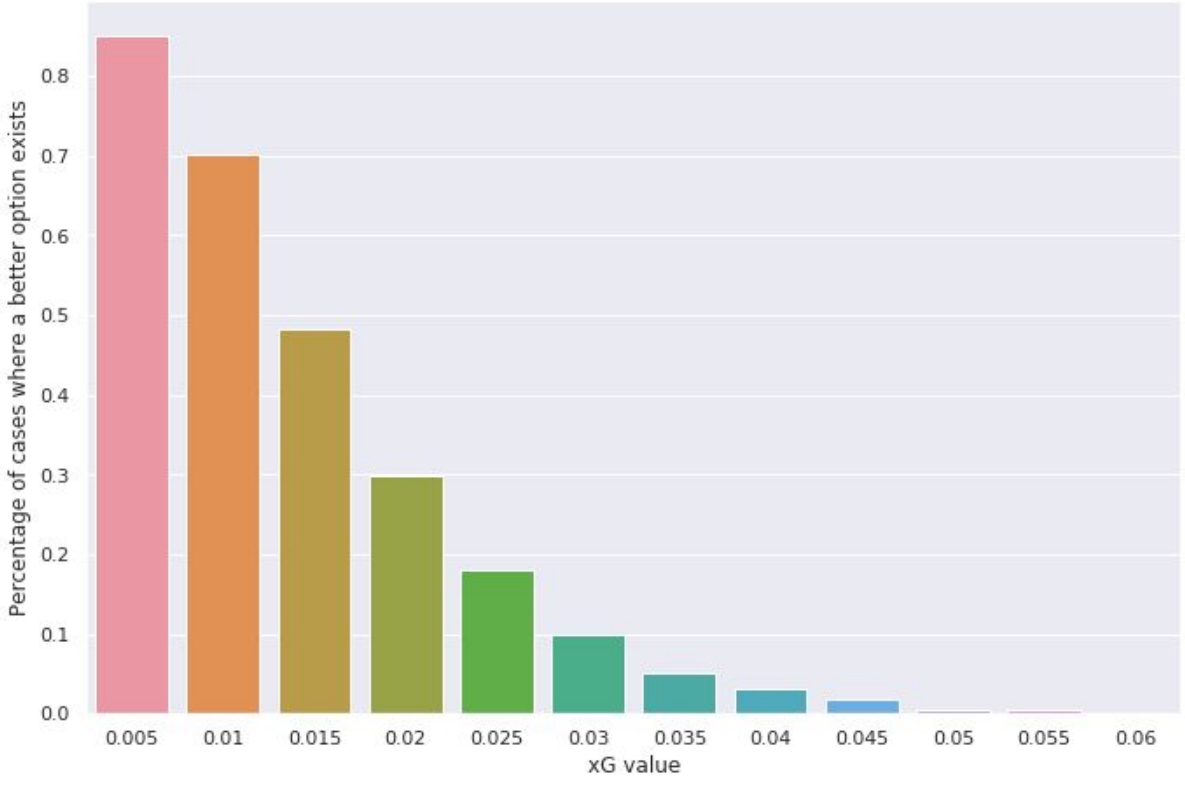

Fig. 7: percentage of cases were at least a better option rather than shooting exists, depending on xG value.

We can clearly see that the percentage of shots with better options increases when the *xG* value decreases. So in addition to knowing that shooting was a poor decision (if the xG value is really small the shooter should have done something else), we know that at least one option was possible and better. We're only showing small *xG* value in Fig. 7 as better options are getting more and more rare, but the volume of small *xG* value shots is big enough to matter anyway.

To illustrate further, we look at percentage of better options with a team perspective. We wonder if a weak percentage of shots with better options (meaning less bad decisions by the team) is correlated with *French Ligue 1* final ranking. In Fig. 8 and 9 we have these percentages by team and final ranking for each club, for seasons 2017/2018 and 2018/2019 respectively.

Team	***% shots with better options***	***Ligue 1 final ranking***
Paris Saint-Germain	14.5	1
Lyon	16.7	3
AS Monaco	17.0	2
Bordeaux	20.3	6
Troyes	22.1	19
OGC Nice	22.2	8
Marseille	22.8	4
Dijon	22.9	11
Metz	23.1	20
Guingamp	23.4	12
Saint Etienne	24.76	7
Angers	24.78	14
Strasbourg	25.6	15
Caen	25.8	16
Toulouse	26.6	18
Lille	27.7	17
Nantes	28.1	9
Rennes	29.7	5
Amiens	30.1	13
Montpellier	31.5	10

Fig. 8: percentage of shots with better options Ligue 1 2017/2018 final ranking.

Team	***% shots with better options***	***Ligue 1 final ranking***
Paris Saint-Germain	17.1	1
Lille	21.1	2
Saint Etienne	21.3	4

Stade de Reims	22.2	8
Lyon	22.4	3
Angers	23.3	13
Marseille	23.6	5
Dijon	23.9	18
Rennes	24.9	10
OGC Nice	24.9	7
Montpellier	25.2	6
As Monaco	25.3	17
Toulouse	26.0	16
Guingamp	26.3	20
Strasbourg	27.1	11
Amiens	28.0	15
Bordeaux	31.3	14
Nîmes	31.6	9
Caen	31.6	19
Nantes	32.3	12

Fig. 9: percentage of shots with better options Ligue 1 2018/2019 final ranking.

As we can see from these tables, there is a relationship between our new metric and team performance. It does not explain everything but combined to other useful metrics it could help to understand the game better.

In addition, we can look at our percentage of better options depending on play pattern, as illustrated in Fig. 10.

Play pattern	***% shots with better options***	***Number of cases***
From Counter	9.4	853
From Goal Kick	20.2	531
From Keeper	21.2	203
Regular play	24	7595

From Kick Off	24.1	166
From Corner	25.3	3057
From Free Kick	25.4	2662
From Throw In	29.3	2965

Fig. 10: percentages of cases were at least a better option than shooting exists, depending on the play pattern of the shot.

We can clearly see that there is less better options from counter, which is not surprising considering that the number of players involved in counter-attack is generally small. As for goalkeeper related play patterns, we can imagine that the defense is well structured and it's harder to have available pass options.

Finally we can note that throw-ins lead to a greater percentage of better options, potentially meaning more bad decisions are made from this play pattern.

4.2 Player analysis

As for player analysis, we can see which players are the best with respect to this metric. This is depicted in Fig. 11. We filtered players with at least 100 shots over two seasons.

Top 20 - Players	***% shots with better options***
Edinson Roberto Cavani	4.9
Jimmy Briand	7.1
Kylian Mbappé Lottin	7.4
Radamel Falcao	8.7
Marcus Thuram	8.8
Neymar da Silva Santos Junior	8.9
Valère Germain	9.6
Júlio Tavares	10.1
Alassane Pléa	10.3

Gaëtan Laborde	11
Emiliano Sala	11.5
Pape Moussa Konaté	14.3
Karl Toko Ekambi	14.4
Stéphane Bahoken	15.1
Houssem Aouar	16.3
Nicolas Pépé	16.9
Wesley Saïd	17.0
Angel Di Maria	17.6
Nolan Roux	18.3
Memphis Depay	19.9

Fig. 11: percentage of cases were at least a better option rather than shooting exists, on a player by player bases.

As we can see, this top 20 is almost exclusively composed of strikers, the exception being *Houssem Aouar*. It's natural because of the plus 100 shots filter. This table gives some indications about how players manage their shots.

For example, *Edison Cavani* has few bad decisions: only 5% of the time he had some other options to consider. As *Cavani* is taking a lot of shots in central areas of the pitch, it i possible that his shots have a high xG value, explaining the low percentage of better options. Nevertheless, it gives some knowledge about shots and how players are managing them.

As an illustration, we also did a comparison on a season by season bases (Fig. 12). For a thorough analysis, It could be possible to do this with a sliding window (sequences of 5 games) or depending on the opposition (result of games, rank of opponent, etc).

Top 20 - Players	***% shots with better options 2017/2018***	***% shots with better options 2018/2019***
Edinson Roberto Cavani	5.0	4.5

Jimmy Briand	5.9	8
Kylian Mbappé Lottin	7.1	7.5
Radamel Falcao	4.8	11.6
Marcus Thuram	7.5	10
Neymar da Silva Santos Junior	4.2	17.5
Valère Germain	8.2	11.1
Júlio Tavares	6	13.6
Alassane Pléa	10.3	Not in ligue 1 anymore
Gaëtan Laborde	3.2	13.8
Emiliano Sala	12.4	8.8
Pape Moussa Konaté	11.1	18.4
Karl Toko Ekambi	14.4	Not in ligue 1 anymore
Stéphane Bahoken	10.3	17.9
Houssem Aouar	17.3	15.5
Nicolas Pépé	18.3	16
Wesley Saïd	18.4	15.8
Angel Di Maria	15.5	19.8
Nolan Roux	17.6	19.5
Memphis Depay	15.9	24.1

Fig. 12: Season by season percentages of cases were at least a better option than shooting exists, Top 20 players.

We can see that some players have almost constant percentages (Cavani, Mbappe), some are getting way worse (Falcao, Neymar, Tavares, Laborde, Konate) and others are getting better (Aouar, Pépé,

Wesley Saïd). It does not tell us what is happening but it's a starting point to study what players have changed in their way of shooting.

5. Conclusion and perspectives

In conclusion, we did a first step toward analyzing shots from the Statsbomb freeze frame data. This new metric, characterizing how often shooting was not the best decision, has a range of applications. Some of them was discuss here: broad shots analysis, relation to xG value, player by player analysis.

It is one of the first use of the freeze frame data, and we are already thinking about some perspectives. Here are some areas of improvement:

- Having more data: it will help for the xG approximation and the pass probability estimation, which would lead to a more precise model.
- Improving the "possible pass" model. For example, it would really help to have body orientation in order to calculate possibles passes more precisely.
- Releasing the teammate hypothesis: teammate which receives the ball could do something else than take a shot, and it would be good to have a way of measuring this.

References

[Ref 1] Lewis, M. (2003). Moneyball: The art of winning an unfair game. New York: W.W. Norton.

[Ref 2] xGoals explanation: https://www.americansocceranalysis.com/explanation

[Ref 3] Spearman, W. (2018). Beyond expected goals. In Proceedings of the 12th MIT sloan sports analytics conference (pp. 1-17).

[Ref 4] Voronoi diagram: https://en.wikipedia.org/wiki/Voronoi_diagram

[Ref 5] Voronoi diagram applied to football: https://medium.com/@Soccermatics/the-geometry-of-attacking-football-bee87e7a749

[Ref 6] Soccer Clubs, Afraid Of Missing Out, Are Joining The Data Revolution Before They're Ready:https://www.forbes.com/sites/robertkidd/2019/06/16/fomo-sees-soccer-clubs-prematurely-join-data-revolution/#7bb9d3d1c5b8